POSOLE: Automated Ontological Annotation for Function Prediction

Karin Verspoor*, Judith Cohn, Susan Mniszewski, Cliff Joslyn
Los Alamos National Laboratory, PO Box 1663, Los Alamos, NM, 87545 USA
*To whom correspondence should be addressed: verspoor@lanl.gov

1. INTRODUCTION

We present the methods utilized in a system aimed at predicting the function of protein targets, as represented by a node in the Gene Ontology (1). The core architecture has been utilized in two distinct contexts: function prediction from text for the BioCreAtIvE evaluation Task 2 (2), and function prediction from protein sequences for the CASP function prediction task (3).

The system we have developed is called POSOLE, or the POSet Ontology Laboratory Environment. POSOLE consists of a set of modules supporting ontology representation, categorization of nodes in the ontology, and analysis. The analysis modules provide support for analysis of the ontological structure, the structure of input queries to the categorization module with respect to that structure, and the structure of the predicted categorization with respect to a given set of expected answers. The system requires the definition of mappers called QueryBuilders for implementation within a specific application. These QueryBuilders define how to map from the relevant input for the application to a set of ontology nodes. For both the BioCreAtIvE and CASP applications, this is done by considering the neighborhood of the protein in the input space and associating entities in the neighborhood to Gene Ontology (GO) nodes. Then POSOLE categorizes the collection of GO nodes based on their distribution in the GO structure, utilizing a technology called POSOC, the POSet Ontology Categorizer (4) (originally called GOC, the Gene Ontology Categorizer (5), but generalized for use with any partially ordered ontology). The resulting set of Gene Ontology nodes is interpreted as the most representative nodes for the function of the input protein. The architecture of the two applications and the common POSOLE modules can be seen in Figure 1.

2. BIOCREATIVE APPLICATION

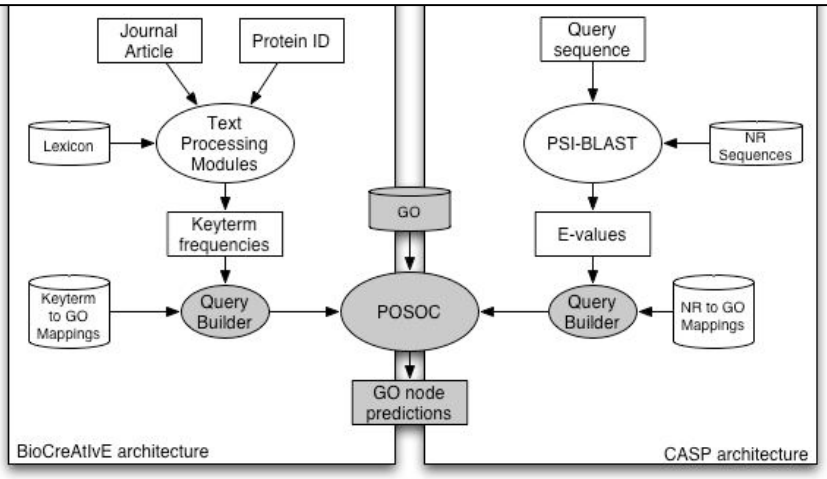
For BioCreAtIvE Task 2, we were provided with a protein identifier (Swiss-Prot identifier) and a relevant journal publication and asked to predict the function of the protein, as represented by a set of GO nodes, on the basis of that publication. The application defines a POSOLE QueryBuilder that is responsible for associating terms in the publication to GO nodes. This is accomplished through the use of natural language processing components. Specifically, the document was processed to morphologically normalize terms to their base forms, identify sentence boundaries, and calculate the relative importance of terms using the statistical measure of TFIDF (term frequency inverse document frequency) with respect to a background corpus. We then identified all references to the input protein in the document and collected the terms in a context window around those references. These terms are considered to be in the contextual neighborhood of the protein, and are assumed to be most indicative of the protein's function. These terms in turn are mapped to specific GO nodes through lexical matches between the text and the text of the GO nodes, in the GO node labels and node definitions as well as in additional sets of terms that were previously associated with specific GO nodes via unsupervised learning (see (2)). An input query for POSOC is constructed which consists of the set of matched GO nodes, weighted according to the TFIDF of the matching term.

3. CASP APPLICATION

For the CASP function prediction task, we were provided with a protein sequence and asked to predict the function of the protein, again in terms of a set of GO nodes. The application defines a POSOLE QueryBuilder that is responsible for associating the input sequence to GO nodes. In this case, we use a "nearest neighbor" approach: we identify close neighbors of the input sequence in sequence space and collect the GO nodes associated with those neighbors in a curated data set (Swiss-Prot).

To identify close neighbors of a target sequence, we performed a PSI-BLAST (Position-Specific Iterated BLAST) (6) search on the target against the NCBI NR database, with 5 iterations. We used the default e-value threshold of 10. Once the nearest neighbors in sequence space of the target sequence have been identified, we collect the GO nodes associated with these sequences. To achieve this, we utilize the

Figure 1: POSOLE application architectures. The architectures for the BioCreAtIvE and CASP protein function prediction applications, built around the core POSOLE modules, shown in grey.



UniProt Swiss-Prot to GO mappings to find all of the nodes related to the corresponding proteins. Finally, we build a weighted collection of GO nodes, where each node in the collection is weighted according to the PSI-BLAST e-value. Several near neighbors of the original target sequence may map to the same nodes, so the collection we build can have redundancy. In this case, each occurrence of a GO node will be weighted individually according to its source.

4. POSOC

For each application, the collection of weighted GO nodes becomes the input query to the categorization technology POSOC (4). This technology aims to identify a set of nodes in a partially ordered set, such as the GO, which best summarize or categorize a given list of input nodes. The technology is based on a view of bio-ontologies as combinatorially structured databases rather than facilities for logical inference, and draws on the discrete mathematics of finite partially ordered sets (*posets*) to develop data representations and algorithms appropriate for the GO. Briefly (for more detail, see (2),(4)), after identifying the set of input nodes in Gene Ontology space, POSOC traverses the structure of the Gene Ontology, percolating hits upwards, and calculating scores for each GO node. POSOC then returns a rank-ordered list of GO nodes representing cluster heads. In the end, this provides an assessment of which nodes best cover the input set. We consider this set of cluster heads to be indicative of the function of the input protein.

5. REFERENCES

1. The Gene Ontology Consortium. 2000. Gene Ontology: Tool For the Unification of Biology, *Nature Genetics*, 25:1:25-29.

2. Verspoor, K., Cohn, J., Joslyn, C., Mniszewski, S., Rechtsteiner, A., Rocha, L.M., Simas, T. 2005. Protein Annotation as Term Categorization into the Gene Ontology using Word Proximity Networks. BMC Bioinformatics 2005 6(suppl 1).

3. Verspoor, K., Cohn, J., Mniszewski, S., Joslyn, C. 2005. Nearest Neighbor Categorization for CASP Function Prediction. In ISMB 2005 poster session. Detroit, MI.

4. http://www.c3.lanl.gov/~joslyn/posoc.html

5. Joslyn, C., Mniszewski, S., Fulmer, A., Heaton, G. (2004). The Gene Ontology Categorizer. Bioinformatics, vol. 20, supplement 1, i169-i177.

6. Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389-3402.